# The ACII 2022 Affective Vocal Bursts Workshop & Competition


Alice Baird
*Hume AI*
New York, USA
alice@hume.ai

Panagiotis Tzirakis
*Hume AI*
New York, USA
panagiotis@hume.ai

Jeffrey A. Brooks
*Hume AI*
New York, USA
jeff@hume.ai

Chris B. Gregory
*Hume AI*
New York, USA
chris@hume.ai

Björn Schuller
*Imperial College London*
London, UK
schuller@ieee.org

Anton Batliner
*University of Augsburg*
Augsburg, Germany
batliner@uni-a.de

Dacher Keltner
*University of California Berkeley*
California, USA
keltner@berkeley.edu

Alan Cowen
*Hume AI*
New York, USA
alan@hume.ai





***Abstract*—The ACII Affective Vocal Bursts Workshop & Competition is focused on understanding multiple affective dimensions of vocal bursts: laughs, gasps, cries, screams, and many other non-linguistic vocalizations central to the expression of emotion and to human communication more generally. This year's competition comprises four tracks using a large-scale and in-the-wild dataset of 59,299 vocalizations from 1,702 speakers. The first, the A-VB-HIGH task, requires competition participants to perform a multi-label regression on a novel model for emotion, utilizing ten classes of richly annotated emotional expression intensities, including; Awe, Fear, and Surprise. The second, the A-VB-TWO task, utilizes the more conventional 2-dimensional model for emotion, arousal, and valence. The third, the A-VB-CULTURE task, requires participants to explore the cultural aspects of the dataset, training native-country dependent models. Finally, for the fourth task, A-VB-TYPE, participants should recognize the type of vocal burst (e.g., laughter, cry, grunt) as an 8-class classification. This paper describes the four tracks and baseline systems, which use state-of-the-art machine learning methods. The baseline performance for each track is obtained by utilizing an end-to-end deep learning model and is as follows: for A-VB-HIGH, a mean (over the 10-dimensions) Concordance Correlation Coefficient (CCC) of 0.5687 CCC is obtained; for A-VB-TWO, a mean (over the 2-dimensions) CCC of 0.5084 is obtained; for A-VB-CULTURE, a mean CCC from the four cultures of 0.4401 is obtained; and for A-VB-TYPE, the baseline Unweighted Average Recall (UAR) from the 8-classes is 0.4172 UAR.**

***Index Terms*—affective computing, vocal bursts, emotional expression, multi-label, machine learning**


## I. INTRODUCTION

The **A**ffective-**V**ocal **B**urst (A-VB ) competition is exploring the expression of affect and emotion in brief nonverbal vocalizations (e. g., vocal bursts such as laughs, sighs, and shouts). Within this competition, the organizers provide several emotion modeling strategies and aim to discuss each during the workshop held at the 2022 Affective Computing and Intelligent Interactions (ACII) Conference.

Thus far, vocal bursts have been largely overlooked in machine learning, affective computing, and emotion science. Given the focus in these fields on facial expressions, the voice has been a relatively understudied medium for communicating emotion. To the extent that the voice has been studied as a modality of emotion expression, it has been chiefly understood from the perspective of speech prosody [1]. But another way humans communicate emotion with the voice is with the brief sounds that occur in the absence of speech – laughs, cries, and shouts (to name a few). Recent studies have discussed the range of emotions conveyed by vocal bursts (known as affect bursts [2], [3]), with findings demonstrating that brief vocalizations reliably express over ten emotions and that the meanings of vocal bursts are generally preserved across diverse cultures [4], [5].

The field of machine learning has recently seen increased interest in vocal burst modeling, with the Expressive Vocalizations (ExVo) competition at ICML in 2022 [6] being the first-of-its-kind competition to explore various modeling strategies to understand and generate vocal bursts. More broadly, computational speech-based emotion modeling has become a prevalent area of research since the success of computational paralinguistic methods [7] and general advances in machine and deep learning speech recognition strategies [8]. Computational modeling of emotion promises to inform a wide range of wellbeing domains, with applications including diagnostic tools for psychiatric illnesses [9], and bio-markers for remote wellness monitoring [10].

In the A-VB competition, we extend on our recent works [6], with a more specific focus on comparing and contrasting the various strategies available for modeling emotion in vocal bursts. In particular, the A-VB competition presents four sub-challenges utilizing a single dataset: (1) the high-dimensional emotion task (A-VB-HIGH), in which participants must predict a high-dimensional (10 class) emotion space, as a multi-output regression task, (2) the two-dimensional emotion task (A-VB-TWO), where the two-dimensional emotion space based on the circumplex model of affect [11] (arousal and valance) is to be recognized, again as a multi-output regression task, (3) the cross-cultural emotion task (A-VB-CULTURE), where participants will be challenged with predicting the intensity of 10 emotions associated with

each vocal burst as a multi-output regression task, using a model or multiple models that generate predictions specific to each of the four cultures provided in the dataset (the U.S., China, Venezuela, or South Africa), and (4) the expressive burst-type task (A-VB-TYPE), in which participants are challenged with classifying the type of expressive vocal burst from 8-classes; *Cry, Gasp, Groan, Grunt, Laugh, Other, Pant, Scream*.

The dataset used within the A-VB competition, the Hume Vocal Bursts dataset (HUME-VB), comprises 59,201 recordings totaling more than 36 hours of audio data from 1,702 speakers. First utilized in the A-VB competition [6], to our knowledge, this dataset remains one of the largest available of vocal bursts. The recordings in HUME-VB are rich and diverse in several ways that present unique opportunities, with the labeling enabling an array of emotion characteristics to be explored from vocal bursts. A single vocal burst can combine classes such as gasps infused with a cry or a scream, ending with a laugh, and offers a vibrant testing bed for emotion understanding and modeling [5]. Thus, the HUME-VB dataset enables distinct but complementary strategies: allowing participants to model continuous blends of utterances such as laughs, cries, and gasps as well as the specific meanings of different laughs (amusement, awkwardness, and triumph), cries (distress, horror, and sadness), gasps (awe, excitement, fear, and surprise), and more.

In this paper, we include a description of the HUME-VB dataset in detail (Section II), provide rules for the four competition tasks (Section III), and present baseline results for each task (Section IV). We summarize our results in Section V and conclude with a discussion of insights from baseline development in Section VI.

## II. THE A-VB DATA

The A-VB competition relies on the HUME-VB dataset, a large-scale dataset of emotional non-linguistic vocalizations (vocal bursts). This dataset consists of 36:47:04 (HH:MM:SS) audio data from 1 702 speakers aged from 20 to 39 years. The data was gathered in 4 countries with broadly differing cultures: China, South Africa, the U.S., and Venezuela, and individuals are performing emotional mimicry of seed emotion examples. Furthermore, speakers' are recorded in their homes via their microphones.

Each vocal burst has been labeled in terms of the intensity of 10 different expressed emotions, each on a [1:100] scale, and these are averaged over an average of 85.2 raters' responses1 *Amusement*2 *Awe*3 *Awkwardness*4 *Distress*5 *Excitement*6 *Fear*7 *Horror*8 *Sadness*9 *Surprise*10 and *Triumph*.

As well as the distribution of arousal and valance, and cultural-based emotion dimensions of HUME-VB, in Figure 1 (left), a t-SNE representation of emotional expressions based on the human ratings across the training set is visualized. We can see that the expressions vary, with clearly defined regions corresponding to each expressed emotion and continuous gradients between emotions (e. g., amusement and excitement).

TABLE I
AN OVERVIEW OF THE HUME-VB COMPETITION DATA. INCLUDING (NO.) SAMPLES, DURATION (HH: MM: SS), SPEAKERS, AND COUNTRY-OF-ORIGIN. THE AGE RANGE FOR SPEAKERS IS 20.5:39.5 YEARS. DUE TO THE COMPETITION SETTING, THE TEST SET FOR THIS DATASET IS BLINDED.

| | **Train** | **Val.** | **Test** | $\sum$ |
|---|---|---|---|---|
| **HH: MM: SS** | 12:19:06 | 12:05:45 | 12:22:12 | 36:47:04 |
| **No.** | 19 990 | 19 396 | 19 815 | 59 201 |
| **Speakers** | 571 | 568 | 563 | 1 702 |
| **USA** | 206 | 206 | — | — |
| **China** | 79 | 76 | — | — |
| **South Africa** | 244 | 244 | — | — |
| **Venezuela** | 42 | 42 | — | — |

Of note, fewer samples convey *Triumph*, so we expect this class to be more challenging to model.

The intensity ratings for each emotion were scaled to [0:1]. For our baseline experiments, the audio files were normalized to -3 decibels and converted to 16 kHz, 16 bit, mono (we also provide participants with the raw unprocessed audio, captured at 48 kHz). The data was subsequently partitioned (see Table I,) into training, validation, and test splits, considering speaker independence and a balance across classes.

## III. THE COMPETITION TASKS

In the A-VB competition, we present four tasks of varying nature utilizing the HUME-VBdata. Each explores a different aspect of the affective samples, with our aim to understand more deeply the various strategies for modeling emotion in vocalizations – an ongoing area of research for machine learning.

### A. A-VB High

In the High-Dimensional Emotion Sub-Challenge (A-VB-HIGH), participants are challenged with predicting the intensity of 10 emotions (Awe, Awkwardness, Amusement, Distress, Excitement, Fear, Horror, Sadness, Surprise, and Triumph) associated with each vocal burst as a multi-output regression task. Participants will report the mean Concordance Correlation Coefficient (CCC) across all ten emotions.

### B. A-VB Two

In the Two-Dimensional Sub-Challenge (A-VB-TWO), participants predict values of arousal and valence (based on 1=unpleasant/subdued, 5=neutral, 9=pleasant/stimulated), derived from the circumplex model for affect [12] as a regression task. Participants will report the mean CCC across the two dimensions. In Figure 1 (middle) we see the distribution of the valence and arousal ratings a t-SNE representation, showing a broad distribution for valence.

### C. A-VB Culture

The Cross-Cultural High-Dimensional Emotion Sub-Challenge (A-VB-CULTURE) is a 10-dimensional, 4-country culture-specific emotion intensity regression task. In A-VB-CULTURE, participants are challenged with predicting the

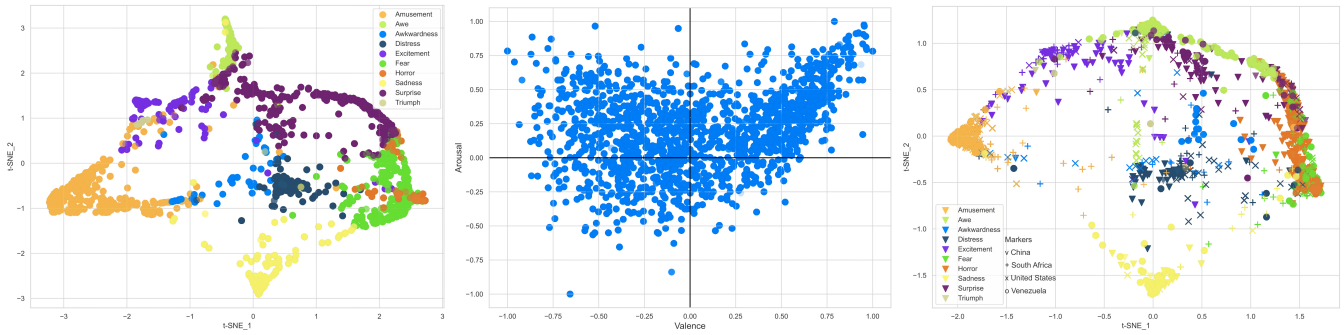


Fig. 1. t-SNE representation of the emotional expression, labeled based on the most dominantly expressed emotion (left), the distribution of arousal and valance (middle), and a t-SNE representation of the culture-based emotion labels (right), from the HUME-VB training set.

intensity of 40 emotions (10 from each culture) as a multi-output regression task. The label for each vocal burst consists of a culture-specific gold standard created from the average of annotations from the sample's culture. Participants will report the mean CCC across all 40 emotions. In Figure 1 (right) a t-SNE representation of the per-culture emotions for each country is plotted, showing clusters of each emotion similar to the A-VB-HIGH task.

### D. A-VB Type

In the Expressive Burst-Type Sub-Challenge (A-VB-TYPE), participants are challenged with classifying the type of expressive vocal burst from 8 classes (Gasp, Laugh, Cry, Scream, Grunt, Groan, Pant, Other). Participants will report the Unweighted Average Recall (UAR) as a measure of accuracy.

### E. General Guidelines

To participate in the A-VB 2022 competition, all participants are asked to provide a completed copy of the HUME-VB End-User License Agreement (EULA) (more details can be found on the competition homepage[1]). Participants should submit a paper that meets the official ACII guidelines, describing their methods. (The A-VB workshop also accepts contributions on related topics.) To obtain test scores, participants should submit their test set predictions to the competition organizers (each team can do this up to 5 times). Participants are free to compete in any or all tasks.

## IV. BASELINE EXPERIMENTS

For each sub-challenge of the A-VB competition, we provide a baseline system utilizing well-established methods known in audio-based emotion recognition modeling [13]–[15]. We provide reproducible code supporting each baseline system on GitHub[2].

### A. Feature-based Approach

We extract two sets of features, each having precedence in related tasks [16]–[18]. One feature vector is extracted per sample for each feature set. Using the OPENSMILE toolkit [19], we extracted the 6,373-dimensional COMPARE set and the 88-dimensional EGEMAPS set. The 2016 COMputational PARalinguistics ChallengE (COMPARE) [20] set contains 6,373 static features computed based on functionals from low-level descriptors (LLDs) [16], [21]. The extended Geneva Minimalistic Acoustic Parameter Set (EGEMAPS) [14], is smaller in size (88-dimensions), and designed for affective-based computational paralinguistic tasks.

We apply a standard neural network (NN) for these experiments. The NN consists of three fully-connected layers, with layer normalization between each and a leaky rectified linear unit (Leaky ReLU) as the activation function. For the regression experiments, sigmoid is applied to the output layer. The loss for each task is varied, with multi-label emotion experiments utilizing a combined Mean Square Error (MSE) loss and the classification tasks applying cross-entropy loss, including softmax on the output layer. From several experiments for each task, a global learning rate ($lr$) and batch size ($bs$) is chosen of $lr = 10^{-3}$ and $bs = 8$. We also apply early stopping (patience of 5, maximum of 25 epochs) to avoid the effects of overfitting the model.

### B. End-to-End Approach

For our end-to-end baseline, we use the multimodal profiling toolkit END2YOU [15]. The baseline model comprises a convolutional neural network (CNN) that extracts features from each audio frame and a recurrent neural network (RNN) that extracts temporal features. We use the Emo-18 (CNN) network architecture [22], which consists of three cascade blocks of 1-D CNN layers, a Leaky ReLU activation function ($\alpha = 0.1$), and max-pooling operations. Both convolution and pooling operations are performed in the time domain, using the raw waveform as input. Before the final emotion prediction, we exploit temporal patterns in the signals using a 2-layer Long-Short Term Memory (LSTM) network.

The input audio frame passed to the CNN is $0.1$ sec long, corresponding to a $1\,600$ dimensional vector, corresponding to the audio sampling rate of 16 kHz. Audio signals with a length not divisible by the input length are padded with zeros.

Our model is trained with the Adam optimization algorithm [23], a batch size of $8$, and an initial learning rate of

[1]http://competitions.hume.ai/avb2022

[2]http://github.com/HumeAI/competitions/tree/main/A-VB2022

TABLE II

BASELINE SCORES FOR A-VB 2022. REPORTING THE MEAN CONCORDANCE CORRELATION COEFFICIENT (CCC) FOR THE THREE REGRESSION TASKS AND THE UNWEIGHTED AVERAGE RECALL (UAR) ACROSS THE 8-CLASSES (CHANCE LEVEL .125) FOR A-VB-TYPE. FOR EACH TASK, THE BEST SCORE ON THE TEST SET IS EMPHASIZED AS THE OFFICIAL BASELINE. WE REPORT THE BEST SCORES FROM 5 SEEDS.

| Approach | CCC | | | | | | UAR | |
|---|---|---|---|---|---|---|---|---|
| | A-VB-HIGH | | A-VB-TWO | | A-VB-CULTURE | | A-VB-TYPE | |
| | Val. | Test | Val. | Test | Val. | Test | Val. | Test |
| COMPARE | .5154 | .5214 | .4942 | .4986 | .3867 | .3887 | .3913 | .3839 |
| EGEMAPS | .4484 | .4496 | .4114 | .4143 | .3229 | .3214 | .3608 | .3546 |
| END2YOU | .5638 | **.5686** | .4988 | **.5084** | .4359 | **.4401** | .4166 | **.4172** |

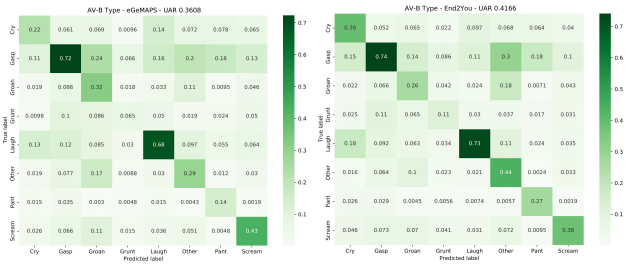


Fig. 2. Normalized confusion matrix for validation results of A-VB-TYPE, with EGEMAPS (left) and END2YOU (right) approaches.

$10^{-4}$. The network weights have been initialized with Kaiming uniform [24] initialization, and the biases are initially set to zero. The LSTM network comprises 256 hidden units and is trained with a gradient norm clipping of 5.0. Finally, we use the MSE loss function and the CCC evaluation metric for the regression tasks. For the classification task, we use the cross-entropy loss with UAR as the evaluation metric.

## V. DISCUSSION OF COMPETITION BASELINES

In Table II, we provide the baseline results for each of the four sub-challenges of the A-VB competition. In all cases, the baseline score is set by the end-to-end approach END2YOU, with feature-based strategies falling short in all cases.

For the A-VB-HIGH task, a baseline on the test set of 0.5687 CCC is obtained utilizing the end-to-end, END2YOU method. Of interest here, we see the COMPARE features closely following much better than EGEMAPS. This suggests that the prosodic- and spectral-based features included with the COMPARE set may benefit this task. On the other hand, the limited samples available may also restrict the potential performance possible from the END2YOU method.

We see similar results for A-VB-TWO, with a baseline on the test set of 0.5084 CCC obtained for the mean across the two classes, arousal, and valance. Of interest, we find that the score for valance is higher than for arousal, 0.5701 and 0.4468 CCC, respectively. Typically, arousal would be easier than valance to model from speech [25]. However, arousal tends to correlate highly with traits including speech-rate [26], and volume [27]. With this in mind, this data is non-language based, and we consider that arousal may be more of a challenge in this context, as these samples are mainly single bursts, and volume may be less impacting on the perception of arousal given varied recording environments.

As with A-VB-HIGH and A-VB-TWO, the baseline is set by the END2YOU approach for A-VB-CULTURE, with a CCC of 0.4401 CCC on the test set. Given the multi-cultural nature of this task, the overall CCC is lower than the others, as some cultures are more difficult to model. This deficiency is shown for Venezuela (a mean of 0.3888 CCC) and China (a mean of 0.3870 CCC), possibly due to the smaller sample size and the cultural difference in these samples.

For the A-VB-TYPE task, we explore classification for the first time with this data, classifying 8-classes of vocalization types. Once again, the END2YOU approach is set as the baseline (0.4172 UAR on the test set), with a similar margin to the hand-crafted feature-based methods. In Figure 2, we can see the confusion matrix for the test results of the baseline system and the EGEMAPS approach. The most commonly confused class appears to be 'Gasp' in both cases, possibly caused by the class imbalance, given that the 'Gasp' class is the most dominant (7,104 samples vs. 4,940 for 'Laugh', on the training set). Furthermore, we see that the hand-crafted features perform better for some classes, mainly 'Screaming' in the case of EGEMAPS; this may indicate that the speech-based features are valuable for this task, supported by their strong performance across tasks.

## VI. CONCLUDING REMARKS

This contribution introduced the guidelines and baseline scores for the first ACII Affective Vocal Bursts (A-VB ) competition. The competition focuses on strategies for computationally modeling emotion in vocal bursts and utilizes a large-scale dataset, the HUME-VB corpus. In this year's A-VB , four tasks are presented: (1) A-VB-HIGH, a multi-label regression task utilizing 10 dimensions of emotion, we report a baseline score of **0.5686 CCC for A-VB-HIGH**; (2) A-VB-TWO, modeling two-dimensions of emotion (arousal and valance), we report a baseline score of **CCC of 0.5084 for A-VB-TWO**; (3) A-VB-CULTURE in which participants should model 40 emotional dimensions, 10 for each culture, we report a baseline score of **0.4401 CCC for A-VB-CULTURE**; and (4) A-VB-TYPE, an 8-class classification of vocalization type, we report a baselines score of **0.4172 UAR for A-VB-TYPE**. Several aspects can be explored by participants of the A-VB competition to improve on the provided baselines. Namely, exploring the advantages of jointly learning from the various labeling provided and knowledge-based approaches which harness the diversity present in the HUME-VB dataset.

## REFERENCES


[1] K. R. Scherer, T. Johnstone, and G. Klasmeyer, "Vocal expression of emotion," *Handbook of affective sciences*, pp. 433–456, 2003.

[2] K. R. Scherer, "Expression of emotion in voice and music," *Journal of voice*, vol. 9, no. 3, pp. 235–248, 1995.

[3] M. Schröder, "Experimental study of affect bursts," *Speech communication*, vol. 40, no. 1-2, pp. 99–116, 2003.

[4] D. T. Cordaro, D. Keltner, S. Tshering, D. Wangchuk, and L. M. Flynn, "The voice conveys emotion in ten globalized cultures and one remote village in bhutan." *Emotion*, vol. 16, no. 1, p. 117, 2016.

[5] A. S. Cowen, H. A. Elfenbein, P. Laukka, and D. Keltner, "Mapping 24 emotions conveyed by brief human vocalization." *American Psychologist*, vol. 74, no. 6, p. 698, 2019.

[6] A. Baird, P. Tzirakis, G. Gidel, M. Jiralerspong, E. B. Muller, K. Mathewson, B. Schuller, E. Cambria, D. Keltner, and A. Cowen, "The ICML 2022 Expressive Vocalizations Workshop and Competition: Recognizing, Generating, and Personalizing Vocal Bursts," *In Proc. The ICML 2022 Expressive Vocalizations (ExVo) Workshop and Competition, arXiv preprint arXiv:2205.01780*, 2022.

[7] B. Schuller and A. Batliner, *Computational Paralinguistics – Emotion, Affect, and Personality in Speech and Language Processing*. Chichester, UK: Wiley, 2014.

[8] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen *et al.*, "Deep speech 2: End-to-end speech recognition in english and mandarin," in *International conference on machine learning*. PMLR, 2016, pp. 173–182.

[9] J. C. Mundt, A. P. Vogel, D. E. Feltner, and W. R. Lenderking, "Vocal acoustic biomarkers of depression severity and treatment response," *Biological psychiatry*, vol. 72, no. 7, pp. 580–587, 2012.

[10] A. Coravos, S. Khozin, and K. D. Mandl, "Developing and adopting safe and effective digital biomarkers to improve patient outcomes," *NPJ digital medicine*, vol. 2, no. 1, pp. 1–5, 2019.

[11] J. A. Russell, "A Circumplex Model of Affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.

[12] ——, "A circumplex model of affect." *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[13] B. Schuller, S. Steidl, A. Batliner, J. Hirschberg, J. K. Burgoon, A. Baird, A. Elkins, Y. Zhang, E. Coutinho, and K. Evanini, "The INTERSPEECH 2016 Computational Paralinguistics Challenge: Deception, Sincerity & Native Language," in *Proc. Interspeech*, San Francisco, CA, 2016, pp. 2001–2005.

[14] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2015.

[15] P. Tzirakis, S. Zafeiriou, and B. W. Schuller, "End2you–the imperial toolkit for multimodal profiling by end-to-end learning," *arXiv preprint arXiv:1802.01115*, 2018.

[16] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The Interspeech 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism," in *Proc. Interspeech*, Lyon, France, 2013, pp. 148–152.

[17] M. Schmitt, F. Ringeval, and B. Schuller, "At the Border of Acoustics and Linguistics: Bag-of-Audio-Words for the Recognition of Emotions in Speech," in *Proc. Interspeech*, San Francisco, CA, 2016, pp. 495–499.

[18] L. Stappen, A. Baird, G. Rizos, P. Tzirakis, X. Du, F. Hafner, L. Schumann, A. Mallol-Ragolta, B. W. Schuller, I. Lefter, E. Cambria, and I. Kompatsiaris, "Muse 2020 challenge and workshop: Multimodal sentiment analysis, emotion-target engagement and trustworthiness detection in real-life media," in *Proceedings of the 1st International on Multimodal Sentiment Analysis in Real-Life Media Challenge and Workshop*. ACM, 2020, p. 35–44.

[19] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM International Conference on Multimedia*, 2010, pp. 1459–1462.

[20] B. Schuller, S. Steidl, A. Batliner, J. Hirschberg, J. K. Burgoon, A. Baird, A. Elkins, Y. Zhang, E. Coutinho, K. Evanini *et al.*, "The interspeech 2016 computational paralinguistics challenge: Deception, sincerity & native language," in *Proceedings of INTERSPEECH*, 2016, pp. 2001–2005.

[21] F. Eyben, F. Weninger, F. Groß, and B. Schuller, "Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor," in *Proc. ACM Multimedia*, Barcelona, Spain, 2013, pp. 835–838.

[22] P. Tzirakis, J. Zhang, and B. W. Schuller, "End-to-end speech emotion recognition using deep neural networks," in *Proc. ICASSP*, 2018, pp. 5089–5093.

[23] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[25] B. W. Schuller, A. Batliner, C. Bergler, E.-M. Messner, A. Hamilton, S. Amiriparian, A. Baird, G. Rizos, M. Schmitt, L. Stappen *et al.*, "The interspeech 2020 computational paralinguistics challenge: Elderly emotion, breathing & masks," *Proc. INTERSPEECH. Shanghai, China: ISCA*, 2020.

[26] M. H. Hecker, K. N. Stevens, G. von Bismarck, and C. E. Williams, "Manifestations of task-induced stress in the acoustic speech signal," *The Journal of the Acoustical Society of America*, vol. 44, no. 4, pp. 993–1001, 1968.

[27] C. Hendrick and D. R. Shaffer, "Effects of arousal and credibility on learning and persuasion," *Psychonomic Science*, vol. 20, no. 4, pp. 241–243, 1970.